3-29-2007

FAST ADAPTIVE PENALIZED SPLINES

Tatyana Krivobokova
Department of Economics, University of Bielefeld, Bielefeld, Germany

Ciprian M. Crainiceanu
Johns Hokins Bloomberg School of Public Health, Department of Biostatistics, ccrainic@jhsph.edu

Goran Kauermann
Department of Economics, University of Bielefeld, Bielefeld, Germany

Fast Adaptive Penalized Splines

Tatyana Krivobokova*
University of Bielefeld, Germany

Ciprian M. Crainiceanu†
Johns Hopkins University

Göran Kauermann
University of Bielefeld, Germany

23rd November 2006

Abstract

This paper proposes a numerically simple routine for locally adaptive smoothing. The locally heterogeneous regression function is modelled as a penalized spline with a smoothly varying smoothing parameter modelled as another penalized spline. This is being formulated as hierarchical mixed model, with spline coefficients following a normal distribution, which by itself has a smooth structure over the variances. The modelling exercise is in line with Baladandayuthapani, Mallick & Carroll (2005) or Crainiceanu, Ruppert & Carroll (2006). But in contrast to these papers Laplace's method is used for estimation based on the marginal likelihood. This is numerically simple and fast and provides satisfactory results quickly. We also extend the idea to spatial smoothing and smoothing in the presence of non normal response.

Keywords: Function of locally varying complexity; Hierarchical mixed model; Laplace approximation

*Department of Economics, University of Bielefeld, Postfach 100131, 33501 Bielefeld, Germany (emails: tkrivobokova@wiwi.uni-bielefeld.de, gkauermann@wiwi.uni-bielefeld.de)

†Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N.Wolfe St. E3636 Baltimore, MD 21205 (email: ccrainic@jhsph.edu)

1 Introduction

The recent last years have seen an increasing use of penalized spline smoothing. Originally introduced by O'Sullivan (1986) it was Eilers & Marx (1996) who gave it the name P-spline smoothing. The idea is quite simple. A smooth unknown regression function is estimated by assuming a functional parametric shape constructed via a high dimensional basis function. The dimension of the basis is thereby chosen in a generous way such that sufficient flexibility is achieved. Instead of simple parametric fitting, however, which would yield a highly variable estimate due to the large dimension of the basis, the basis coefficients are penalized such that the resulting fit is smooth. The idea of P-spline has led to a powerful and applicable smoothing technique which is well demonstrated and motivated in the book by Ruppert, Wand & Carroll (2003). The actual dimension of the basis used has thereby little influence on the fit as has been shown in Ruppert (2002) who concludes that "at most 35 to 40 knots (means basis functions) could be recommended for all sample sizes and for all smooth functions without too many oscillations".

The idea of P-spline smoothing can be linked to mixed models as shown in Wand (2003). This particularly allows the use of mixed models software for smoothing and in fact the P-spline fit is equivalent to a Best Linear Unbiased Predictor in the mixed model formulation. In turn, the smoothing the penalty parameter plays the role of the ratio of the random effect variance and residual variance in the mixed model formulation. This allows for smoothing parameter selection with mixed models technology (see Kauermann, 2004). Accordingly, software for fitting penalized splines can take advantage of the affinity to mixed models and its Bayesian formulation as well, see Ngo & Wand (2004), Crainiceanu, Ruppert & Wand (2005) or Lang & Brezger (2004).

Even though P-spline smoothing is easy and practical (see Wand, 2003), the standard setting with a single penalization parameter fails if the function to be estimated is locally of varying complexity, that is if the function is changing rapidly in some regions while in other regions the function is very smooth. This is the general problem of spatially adaptive smoothing which has been treated by a number of authors. For kernel based methods Fan & Gijbels (1995) or Herrmann (1997) may serve as references. For spline smoothing Luo & Wahba (1997) suggest what they call hybrid adaptive splines. The idea is to replace the n dimensional spline basis, where n is the sample size, by a subset of the basis functions with the spline basis functions chosen adaptively. This idea has similarities to adaptive knot selection for regression splines as suggested in Friedman & Silverman (1989). An alternative approach is to allow the smoothing parameter to vary locally adaptive. Using a reproducing Hilbert space formulation this has been suggested in Pintore, Speckman & Holmes (2005) using piecewise constant smoothing parameters. Similarly, making use of the P-spline idea, as also discussed in this paper, Ruppert & Carroll (2000) allow the penalty to act differently for each locally defined spline basis, where the smoothing parameters are then selected using a multivariate generalized cross validation. A similar approach is suggested in Wood, Jiang & Tanner (2002) working with mixtures of splines in a fully Bayesian framework.

In this paper, we stick to the P-spline approach in the line of Ruppert & Carroll (2000) and achieve spatial adaptivity by imposing a functional structure on the smoothing parameters. This is in line with Baladandayuthapani, Mallick & Carroll (2005) and Crainiceanu, Ruppert & Carroll (2006) who additionally allow for local heterogeneity. Lang & Brezger (2004) achieve a local adaptive P-spline by pursuing a Bayesian model of P-splines where spline coefficients trace from a heterogenous

random walk. Due to the Bayesian framework the latter papers require the use of MCMC methods to obtain an estimate. The intention of this paper is to demonstrate how the MCMC techniques can be easily circumvented by simple Laplace approximation. Even though this is a step back in terms of the technical features we have nowadays, it is a step forward in terms of simplicity of numerics and therewith allowing for fast calculation.

The paper is organized as follows. In Section 2 we introduce our spatially adaptive modelling which is evaluated by simulations. Section 3 extends the results to spatial smoothing, again including simulations. Section 4 generalizes approach to non-normal response case following by simulations and an example of adaptive bivariate smoothing of binary data. A conclusion finishes the paper.

2 Smoothly varying local penalties for P-spline regression

2.1 Hierarchical penalty model

Our model is

$$y_i \sim N(m(x_i), \sigma_\epsilon^2), \quad i = 1, ..., n, \tag{1}$$

where $m(x)$ is a smooth function in the univariate metrical quantity x. We assume that $m(x)$ can be of locally varying complexity and replace $m(x)$ for fitting by the penalized truncated polynomials

$$m(x) = \beta_0 + x\beta_1 + ... + x^q\beta_q + \sum_{s=1}^{K_b}(x - \tau_s^{(b)})_+^q b_s, \tag{2}$$

where $\tau_1^{(b)}, ..., \tau_{K_b}^{(b)}$ are knots covering the range of x and $(x - \tau_s^{(b)})_+^q$ is the truncated q-th order polynomial defined through $(x - \tau_s^{(b)})^q$ if $x - \tau_s^{(b)} > 0$ and zero otherwise. The dimension K_b of the basis is chosen in a lush and generous manner and knots

$\tau_s^{(b)}$ are placed over the range of x, e.g. using the quantiles of x. In practice we follow the guideline suggested by Ruppert (2002) and set $K_b \geq \min(n/4, 40)$. Instead of fitting model (2) directly to the data one imposes a penalty on coefficients b_s to achieve a smooth fit. A conventional approach is to penalize $b = (b_1, ..., b_{K_b})$ by the quadratic form $\lambda b^T D b$ with λ as penalization parameter and D as penalty matrix chosen according to the data. For truncated polynomial it has been found useful to choose D as the identity matrix, that is the penalty takes the form $\lambda b^T b$. If instead a B-spline basis is used, the conventional penalty used is constructed from differences between neighbouring spline coefficients (see Eilers & Marx, 1996). Both approaches are in fact closely linked (see Ruppert, Wand & Carroll, 2003). In general, the approach presented is not restricted to truncated polynomials and we use different basis function in our examples subsequently. For simplicity of notation we present our routine for truncated polynomials, without loss of generality though. The interesting feature of spline smoothing is its link to linear mixed models, simply by formulating the penalty as *a priori* distribution on the spline coefficients. This means we model $b \sim N(0, \sigma_b^2 D^-)$ where $\sigma_b^2 = \sigma_\epsilon^2/\lambda$ and D^- as (generalized) inverse of D. For truncated polynomials we chose $D = I$ so that $b \sim N(0, \sigma_b^2 I)$. The restriction explicitly occurring with this setting is that all coefficients have the same *a priori* variance and therewith undergo the same penalization. This is a critical point if the underlying function is of locally varying complexity. Like Crainiceanu, Ruppert & Carroll (2006) or Baladandayuthapani, Mallick & Carroll (2005) we therefore allow coefficients $b_1, ..., b_{K_b}$ to have locally varying variability which is accommodated by

$$b_s \sim N(0, \sigma_{bs}^2), \quad s = 1, ..., K_b.$$

We assume next that the variance components σ_{bs}^2 change smoothly over the (ordered) spline coefficients, meaning that the complexity of function $m(x)$ varies

smoothly over x and does not change rapidly. A typical example for such function is the Doppler curve (see Figure 1). We accommodate this assumption by setting $\sigma^2_{bs} = \sigma^2_b(\tau^{(b)}_s)$, where $\sigma^2_b(.)$ is a function smoothly varying over the knots of the basis. In a hierarchical manner the smooth structure is again modelled by P-splines. To do so we set

$$\sigma^2_b(\tau^{(b)}) = \exp[\gamma_0 + \tau^{(b)}\gamma_1 + ... + \tau^{(b)^p}\gamma_p + \sum_{t=1}^{K_c}(\tau^{(b)} - \tau^{(c)}_t)^p_+ c_t], \tag{3}$$

where $\tau^{(c)}_1, ..., \tau^{(c)}_{K_c}$ is a second layer of knots covering the range of $\tau^{(b)}_1, \ldots, \tau^{(b)}_{K_b}$. Note, that $K_c < K_b$ is a restriction to be held and practically K_c is chosen far smaller than K_b. Extending now the smooth estimation we fit $\sigma^2_b(\cdot)$ in a penalized form by imposing a penalty on coefficients c_t. From a Bayesian viewpoint this can be expressed as *a priori* distribution in the form

$$c_t \sim N(0, \sigma^2_c), \quad t = 1, ..., K_c.$$

Note, that the variance σ^2_c is set to be constant and serves as hyper parameter in our model construction.

For notational simplicity we rewrite the model in matrix form. Let therefore

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X_b = \begin{pmatrix} 1 & \ldots & x^q_1 \\ \vdots & & \vdots \\ 1 & \ldots & x^q_n \end{pmatrix}, \quad Z_b = \begin{pmatrix} (x_1 - \tau^{(b)}_1)^q_+ & \ldots & (x_1 - \tau^{(b)}_{K_b})^q_+ \\ \vdots & & \vdots \\ (x_n - \tau^{(b)}_1)^q_+ & \ldots & (x_n - \tau^{(b)}_{K_b})^q_+ \end{pmatrix}$$

and write $\beta = (\beta_0, \ldots, \beta_q)^T$ and $b = (b_1, \ldots, b_{K_b})^T$. In analogous way we define

$$X_c = \begin{pmatrix} 1 & \tau^{(b)}_1 & \ldots & \tau^{(b)^p}_1 \\ \vdots & & & \vdots \\ 1 & \tau^{(b)}_{K_b} & \ldots & \tau^{(b)^p}_{K_b} \end{pmatrix}, \quad Z_c = \begin{pmatrix} (\tau^{(b)}_1 - \tau^{(c)}_1)^p_+ & \ldots & (\tau^{(b)}_1 - \tau^{(c)}_{K_c})^p_+ \\ \vdots & & \vdots \\ (\tau^{(b)}_{K_b} - \tau^{(c)}_1)^p_+ & \ldots & (\tau^{(b)}_{K_b} - \tau^{(c)}_{K_c})^p_+ \end{pmatrix}$$

which gives the hierarchical model

$$\begin{aligned} Y|b,c &= X_b\beta + Z_b b + \epsilon,\ \epsilon \sim N(0,\sigma_\epsilon^2 I_n), \\ b|c &\sim N(0,\Sigma_b),\ \Sigma_b = \mathrm{diag}[\exp(X_c\gamma + Z_c c)], \\ c &\sim N(0,\sigma_c^2 I_{K_c}). \end{aligned} \tag{4}$$

The corresponding likelihood results to

$$\begin{aligned} L(\beta,\gamma,\sigma_\epsilon^2,\sigma_c^2) &= f(Y;\beta,\gamma,\sigma_\epsilon^2,\sigma_c^2) \\ &= (2\pi)^{-\frac{(n+K_c)}{2}}\sigma_\epsilon^{-n}\sigma_c^{-K_c}\int_{R^{K_c}}\exp[-g(c)]dc, \end{aligned} \tag{5}$$

with

$$g(c) = \frac{1}{2}\log|V_\epsilon| + \frac{c^T c}{2\sigma_c^2} + \frac{(Y-X_b\beta)^T V_\epsilon^{-1}(Y-X_b\beta)}{2\sigma_\epsilon^2}$$

and $V_\epsilon = I_n + Z_b\Sigma_b Z_b^T/\sigma_\epsilon^2$. Note that both, V_ϵ as well as Σ_b, depend on c and γ which is omitted throughout the paper for notational simplicity. The integral in (5) is not available analytically, which motivates a solution based on MCMC techniques as pursued in the previously cited papers. We however go a different route via Laplace approximation, which is justifiable for two reasons. First, the hierarchical model (4) is used as a vehicle for estimation only and has no specific data generating justification. This means finding the exact marginal likelihood by extensive numerics is not necessary, if an approximate version fulfils the task of estimation properly. Secondly, since K_c (and K_b) are assumed to be bounded while sample size n is growing, i.e. $K_c < K_b \ll n$, one finds function $g(\cdot)$ to be of order n. This implies that the Laplace approximation has an error of order $O(n^{-1})$ (see Severini, 2000). Therefore the Laplace approximation appears as attractive alternative to simulation

based techniques. The log-likelihood is then approximated, up to a constant, by

$$\begin{aligned} -2l(\beta,\gamma,\sigma_\epsilon^2,\sigma_c^2) &\approx n\log\sigma_\epsilon^2 + K_c\log\sigma_c^2 + \log|V_\epsilon(\hat{c})| + \log|I_{cc}(\hat{c})| \\ &+ \hat{c}^T\hat{c}/\sigma_c^2 + (Y - X_b\beta)^T V_\epsilon^{-1}(\hat{c})(Y - X_b\beta)/\sigma_\epsilon^2, \end{aligned} \tag{6}$$

where $\hat{c}_t,\ t = 1, ..., K_c$ is the solution to

$$\frac{\partial g(\hat{c})}{\partial c_i} = \frac{1}{2}\mathrm{tr}\left(V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_i}\right) + \frac{c_i}{\sigma_c^2} - \frac{1}{2\sigma_\epsilon^2}(Y - X_b\beta)^T V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_i}V_\epsilon^{-1}(Y - X_b\beta) = 0 \tag{7}$$

and

$$(I_{cc}(c))_{ij} = E\left(\frac{\partial^2 g(c)}{\partial c_i \partial c_j}\middle| c\right) = \frac{\delta_{ij}}{\sigma_c^2} + \frac{1}{2}\mathrm{tr}\left(V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_i}V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_j}\right), \tag{8}$$

with δ_{ij} as the Kronecker delta. It is not difficult to see that the derivative appearing in the above equations results to

$$\frac{\partial V_\epsilon}{\partial c_i} = Z_b\mathrm{diag}(Z_{c,i})\Sigma_b Z_b^T/\sigma_\epsilon^2,$$

where $Z_{c,i}$ stands for the ith column of the matrix Z_c. Moreover, noting that the prediction of b is defined through

$$Z_b^T V_\epsilon^{-1}(y - X_b\beta) = \sigma_\epsilon^2\Sigma_b^{-1}\hat{b}.$$

and $\mathrm{tr}\,(V_\epsilon^{-1}\partial V_\epsilon/\partial c) = Z_c^T w_{df}$, with w_{df} as K_b dimensional vector containing the diagonal elements of $A = Z_b^T Z_b(\sigma_\epsilon^2\Sigma_b^{-1} + Z_b^T Z_b)^{-1}$, we can represent (7) and (8) as

$$\frac{\partial g(c)}{\partial c} = -\frac{1}{2}Z_c^T\left\{\Sigma_b^{-1}\hat{b}^2 - w_{df}\right\} + \frac{c}{\sigma_c^2} = 0,$$

and

$$I_{cc}(c) = E\left(\frac{\partial^2 g(c)}{\partial c \partial c^T}\middle| c\right) = \frac{1}{2}Z_c^T\mathrm{diag}(v_{df})Z_c + \frac{I_{K_c}}{\sigma_c^2},$$

with v_{df} as K_b dimensional vector containing the diagonal elements of AA. Note that $df_b = \sum_{s=1}^{K_b} w_{df} = 1_{K_b}^T w_{df}$ measures the degree of freedom used for fitting b. In

particular, for K_b assumed to be fixed we find $df_b \to K_b$ as n tends to infinity and both w_{df} and v_{df} tend to 1_{K_b}.

Assuming that weights v_{df} vary slowly or not at all as a function of γ (which is readily seen from $\partial v_{df}/\partial\gamma_i = 2\text{diag}[(AA - AAA)\text{diag}(X_{c,i})]$ with $X_{c,i}$ as the ith column of the matrix X_c) we can estimate γ and c simultaneously, resulting in the following iterated weighted least squares (IWLS) for estimation of parameter $\theta = (\gamma^T, c^T)^T$

$$\hat{\theta} = \left(W_c^T \text{diag}\left(\frac{v_{df}}{2}\right) W_c + \frac{D_c}{\sigma_c^2} \right)^{-1} W_c^T \text{diag}\left(\frac{v_{df}}{2}\right) u, \tag{9}$$

with $W_c = (X_c, Z_c)$, $D_c = \text{diag}(0_{(p+1)\times(p+1)}, I_{K_c})$ and $u = W_c\theta + \text{diag}(v_{df}^{-1})(\Sigma_b^{-1}\hat{b}^2 - w_{df})$ as a working vector. Fixing now parameter $\hat{\theta}$ provides, with the above log-likelihood (6), the following parameter estimates

$$\begin{aligned} \hat{\sigma}_c^2 &= \hat{c}^T\hat{c}/w_{df}^c \\ \hat{\beta} &= (X_b^T V_\epsilon^{-1}(\hat{\theta}) X_b)^{-1} X_b^T V_\epsilon^{-1}(\hat{\theta}) y, \\ \hat{\sigma}_\epsilon^2 &= (y - X_b\hat{\beta})^T V_\epsilon^{-1}(\hat{\theta})(y - X_b\hat{\beta})/n, \end{aligned} \tag{10}$$

with $w_{df}^c = \text{tr}(Z_c\text{diag}(v_{df})Z_c^T I_{cc}^{-1}/2)$ and obvious definition for $V_\epsilon(\theta)$. Finally, we obtain the estimated best linear unbiased predictor (EBLUP) via

$$\hat{b} = \hat{\Sigma}_b Z_b^T \hat{V}_\epsilon^{-1}(y - X_b\hat{\beta})/\hat{\sigma}_\epsilon^2.$$

The latter steps are standard and available from linear mixed models technology. Estimation can now be carried out with the standard in mixed models framework EM type algorithm (see e.g. Searle, Casella, & McCulloch, 1992 or Breslow & Clayton, 1993) by iterating between (9) and (10) until convergence. It should be noted that the estimation consists of two simple steps and is, therefore, numerically very fast. In fact, for a reasonably sized data set ($n = 1000$) the fit is achieved within seconds using up to date computers, which contrasts the routine from any Monte

Carlo simulation based methods (Crainiceanu, Ruppert & Carroll, 2006 or Baladandayuthapani, Mallick & Carroll, 2005), where fits are available within minutes, but not seconds.

2.2 Restricted maximum likelihood

The above results are presented for maximum likelihood estimates. The use of restricted maximum likelihood (REML) is, however, more common in mixed models (see Harville, 1977). The restricted maximum log-likelihood for the model (4) takes the form

$$l_R(\beta, \gamma, \sigma_\epsilon^2, \sigma_c^2) = l(\beta, \gamma, \sigma_\epsilon^2, \sigma_c^2) - \frac{1}{2} \log |X_b^T V_\epsilon^{-1}(\hat{c}) X_b / \sigma_\epsilon^2|,$$

with $l(\beta, \gamma, \sigma_\epsilon^2, \sigma_c^2)$ as given in (6). The further estimation procedure is identical to that described in the previous section, with the matrix A in w_{df} and v_{df} being replaced by

$$A_R = A - Z_b^T V_\epsilon^{-1} X_b (X_b^T V_\epsilon^{-1} X_b)^{-1} X_b^T Z_b (Z_b^T Z_b + \Sigma_b^{-1} \sigma_\epsilon^2)^{-1}$$

and the variance estimate defined as $\hat{\sigma}_\epsilon^2 = (y - X_b\hat{\beta})^T V_\epsilon^{-1}(\hat{\theta})(y - X_b\hat{\beta})/n - q - 1$. We have compared the performance of both procedures and found little difference in estimates.

2.3 Variance estimation

We denote with $\tilde{m}(x)|c = X_b\tilde{\beta} + Z_b\tilde{b}|c$ the best linear unbiased predictor (BLUP) of the function $m(x)|c = X_b\beta + Z_b b|c$, where $\tilde{\beta} = (X_b^T V_\epsilon^{-1} X_b)^{-1} X_b^T V_\epsilon^{-1} y$ and $\tilde{b}|c = \Sigma_b Z_b^T V_\epsilon^{-1}(y - X_b\tilde{\beta})/\sigma_\epsilon^2$. Note that within the linear mixed model framework the function $m(x)|c$ is random due to randomness of parameter b. Since $\tilde{m}(x)|c$ is unbiased for $m(x)|c$, the confidence intervals for $m(x)|c$ can be obtained from

$$[\tilde{m}(x) - m(x)]|c \sim N(0, \mathrm{Var}[\tilde{m}(x) - m(x)|c]),$$

where $\text{Var}[\tilde{m}(x) - m(x)|c] = \sigma_\epsilon^2 S(\theta) = \sigma_\epsilon^2 W_b (W_b^T W_b + \sigma_\epsilon^2 D_b(\theta))^{-1} W_b^T$ with $W_b = (X_b, Z_b)$ and $D_b(\theta) = \text{diag}(0_{(q+1)\times(q+1)}, \Sigma_b^{-1})$. Using the delta method and unbiasedness of $\tilde{m}(x)|c$ one can approximate the unconditional variance with

$$\text{Var}[\tilde{m}(x) - m(x)] = E[\text{Var}(\tilde{m}(x) - m(x)|c)] + \text{Var}[E(\tilde{m}(x) - m(x)|c)] \approx \sigma_\epsilon^2 S(\hat{c}).$$

Let now $\hat{m}(x)|c = X_b\hat{\beta} + Z_b\hat{b}|c$ denote the estimated best linear unbiased predictor (EBLUP), obtained from $\tilde{m}(x)|c$ by plugging in the estimates of variance parameters. This can be used to obtain a plug in estimate $\widehat{\text{Var}}[\hat{m}(x) - m(x)] \approx \hat{\sigma}_\epsilon^2 S(\hat{\theta})$.
The variance estimate can also be calculated and justified within the Bayesian framework. Assuming parameters $\Sigma_b = \text{diag}[\exp(W_c\theta)]$ and σ_ϵ^2 are known, the posterior distribution of $m(x)$ is $N(\hat{m}(x,\theta), \sigma_\epsilon^2 S(\theta))$, where $\hat{m}(x,\theta) = S(\theta)y$. An empirical Bayes approach would now replace the unknown values Σ_b and σ_ϵ^2 in the prior by estimates and then treat these parameters as if they were known and given in advance. Thus, the approximate posterior distribution of $m(x)$ results in $N(\hat{m}(x,\hat{\theta}), \hat{\sigma}_\epsilon^2 S(\hat{\theta}))$, yielding the same confidence intervals as in the linear mixed model framework.
Even though the variance formula has the advantage of being simple it does not, however, account for the extra variability due to estimation of θ, that is the local varying penalty. This is the price to pay when using Laplace's method instead of a full Bayesian approach. For further discussion we refer to Morris (1983), Laird & Louis (1987), Kass & Steffey (1989) or Ruppert & Carroll (2000). To correct for this we now estimate the posterior variance of $m(x)$ calculated from the joint posterior distribution of b and θ. We, therefore, use the delta-method correction from Kass & Steffey (1989) and obtain

$$\begin{aligned}\text{Var}(m(x)|y) &= E[\text{Var}(\hat{m}(x)|\hat{\theta}, y)] + \text{Var}[E(\hat{m}(x)|\hat{\theta}, y)] \\ &\approx \hat{\sigma}_\epsilon^2 S(\hat{\theta}) + \left(\left.\frac{\partial \hat{m}(x,\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}}\right)^T \text{Var}(\hat{\theta}) \left(\left.\frac{\partial \hat{m}(x,\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}}\right).\end{aligned}$$

As estimate of $\mathrm{Var}(\hat{\theta})$ one can use the inverse of the Fisher matrix $I_{\theta\theta}(\hat{\theta})$ resulting from the last iteration as by-product. The derivative in the last term, ignoring the dependence of $\hat{\sigma}_\epsilon^2$ on θ, results in

$$\left.\frac{\partial \hat{m}(x,\theta)}{\partial \theta_i}\right|_{\theta_i=\hat{\theta}_i} = \hat{\sigma}_\epsilon^2 W_b (W_b^T W_b + \hat{\sigma}_\epsilon^2 \hat{D}_b)^{-1} \tilde{W}_{c,i} \hat{D}_b (W_b^T W_b + \hat{\sigma}_\epsilon^2 \hat{D}_b)^{-1} W_b^T y,$$

with $\hat{D}_b = D_b(\hat{\theta})$ and $\tilde{W}_{c,i} = \mathrm{diag}(0_{(q+1)\times(q+1)}, W_{c,i})$, where $W_{c,i}$ stands for the i-th column of matrix W_c.

2.4 Numerical implementation

For the numerical implementation one can make use of any standard mixed models software. More precisely, we use the following algorithm:

1. Obtain initial estimates for all parameters from a non-adaptive fit, using any mixed model software;
2. Get next estimates for $\hat{\theta}$ and $\hat{\sigma}_c^2$ from (9) and (10);
3. Update estimates for the remaining parameters with a mixed model software, taking the estimated variance matrix $\hat{\Sigma}_b = \mathrm{diag}[\exp(W_c\hat{\theta})]$ into account;
4. Iterate between 2 and 3 until convergence.

We implemented this algorithm in the package "AdaptFit" described below. With respect to the splines we experimented with a number of spline basis functions, such as B-splines of different degree and penalty order, quadratic and cubic truncated polynomials as well as cubic thin plate splines. Although all basis functions produced very similar, in fact almost indistinguishable, results, the cubic thin plate splines demonstrated a slightly better numerical stability and were preferred for the simulation study. Knots dimensions K_b and K_c need also to be chosen carefully to ensure capturing a complex function structure in the regions of a higher variability.

2.5 Simulations and comparisons with other univariate smoothers

We performed a number of simulations. A particular focus is to compare our results with those reported in Ruppert & Carroll (2000) and Baladandayuthapani, Mallick & Carroll (2005). First, for $n = 400$ x equally spaced on $[0, 1]$ and independent $\epsilon_i \sim N(0, 0.2^2)$ we examined the regression function

$$m_1(x) = \sqrt{x(1-x)} \sin\left(\frac{2\pi(1+2^{(9-4j)/5})}{x+2^{(9-4j)/5}}\right),$$

with $j = 6$. We performed 500 simulations with $K_b = 80$ and $K_c = 20$. An exemplary fit (bold) together with confidence intervals (dashed) is shown on Figure 1. The corresponding estimated variance of random effects is shown in Figure 2. Figure 3 displays the pointwise Mean squared error $E(\{\hat{f}(x) - f(x)\}^2)$ with the expectation being replaced by the mean of the simulations. For better visual impression we show a simple smoother (thick line) for the latter. The average MSE over all x's (AMSE) equals 0.0034, which is comparable with 0.0027 reported in Baladandayuthapani, Mallick & Carroll (2005) and 0.0026 of Ruppert & Carroll (2000). We computed also the coverage probabilities of the 95% confidence intervals over all 500 simulated datasets. Figure 4 shows smoothed pointwise coverage probabilities. For small values of $x \leq 0.1$, i.e. in the region with low signal-to-noise ratio, there is clear undercoverage but beyond 0.1, say the coverage probability exceeds 95% being slightly conservative. The average coverage probability results to 94.95%.

Next, we considered the heterogeneous regression function

$$m_2(x) = \exp(-400(x-0.6)^2) + \frac{5}{3}\exp(-500(x-0.75)^2) + 2\exp(-500(x-0.9)^2).$$

Now $n = 1000$ x values are equally spaced on $[0, 1]$ and $\epsilon_i \sim N(0, 0.5^2)$. We applied our approach to 500 simulated datasets, using $K_b = 40$ and $K_c = 4$. Figures 5 and

6 represent one of the simulated fits and estimated variance of random effects correspondingly. The pointwise MSE is shown in Figure 7. The resulted AMSE is equal 0.0048, which is somewhat smaller than 0.0061 and 0.0065, obtained by Baladandayuthapani, Mallick & Carroll (2005) and Ruppert & Carroll (2000) respectively. The smoothed pointwise coverage probabilities can be seen on Figure 8. The average coverage probability for this function equals 95.94%, which is comparable with 95.22% and 96.28% reported by Baladandayuthapani, Mallick & Carroll (2005) and Ruppert & Carroll (2000) correspondingly. For the same setting Baladandayuthapani, Mallick & Carroll (2005) reported also the simulation results for the BARS approach of DiMatteo, Genovese & Kass (2001). BARS employs free-knots splines with the random number and location of knots, using reversible jump MCMC for estimation. The AMSE based on this approach is 0.0043, while the average coverage probability is 94.72%, which is again comparable with our approach.

To demonstrate insensitivity of our approach to the choice of number of subknots K_c we run additionally simulations for the functions $m_1(x)$ and $m_2(x)$ with different values of K_c. AMSE based on 500 simulations for the function $m_1(x)$ using 10, 20 and 30 subknots, respectively, resulted in 0.00344025, 0.00344029 and 0.00344023. AMSE based on 500 simulations for the function $m_2(x)$ based on K_c equal to 4, 10 and 15, respectively, equal to 0.0048405, 0.0048330 and 0.0048313. In general there should be enough subknots to capture the structure of the variance of random effects and further increase of K_c has little effect on the fit.

Our approach also remains stable even if the function does not require adaptive smoothing. The variance of the random effects will be estimated to be nearly constant, having little effect on the resulting fit. To demonstrate this we run a small simulation study based on the function $\sin(2\pi x)$. We used $n = 400$ points and

$\epsilon_i \sim N(0, 0.3^2)$. The AMSE values based on 150 simulations result in 0.0017038 and 0.0017351 for adaptive and non-adaptive estimates, respectively, suggesting that both approaches deliver nearly indistinguishable results.

Overall, our method provides comparable results to other approaches, but with significantly less numerical effort.

3 Spatial smoothing

3.1 Hierarchical modelling

We now generalize the ideas of the previous section to spatial smoothing

$$y_i \sim N(m(\boldsymbol{x}_i), \sigma_\epsilon^2), \quad i = 1, ..., n,$$

with $\boldsymbol{x}_i \in R^2$ and $m(.)$ as a smooth function of 2 covariates. Following Crainiceanu, Ruppert & Carroll (2006) we use radial basis functions (for details see Ruppert, Wand & Carroll, 2003) and choose K_b knots $\boldsymbol{\tau}_1^{(b)}, ..., \boldsymbol{\tau}_{K_b}^{(b)} \in R^2$. This defines the model matrices X_b with i-th row $[1, \boldsymbol{x}_i^T]_{1 \le i \le n}$ while the basis equals $Z_b = Z_{K_b}\Omega_{K_b}^{-1/2}$ where $Z_{K_b} = [\|\boldsymbol{x}_i - \boldsymbol{\tau}_s^{(b)}\|^2 \log \|\boldsymbol{x}_i - \boldsymbol{\tau}_s^{(b)}\|]_{1 \le s \le K_b, 1 \le i \le n}$ and $\Omega_{K_b} = [\|\boldsymbol{\tau}_t^{(b)} - \boldsymbol{\tau}_s^{(b)}\|^2 \log \|\boldsymbol{\tau}_t^{(b)} - \boldsymbol{\tau}_s^{(b)}\|]_{1 \le s,t \le K_b}$ with $\|.\|$ denoting the Euklidean norm in R^2. Including penalties and using the link to linear mixed model we get

$$\begin{aligned} Y|b &= X_b\beta + Z_b b + \epsilon,\ \epsilon \sim N(0, \sigma_\epsilon^2 I_n), \\ b &\sim N(0, \Sigma_b). \end{aligned} \tag{11}$$

Local adaptive smoothing is now implemented by allowing coefficients b to have locally varying variablity. Like above we set subknots $\boldsymbol{\tau}_1^{(c)}, ..., \boldsymbol{\tau}_{K_c}^{(c)} \in R^2$, $K_c < K_b$ and define matrices X_c and Z_c similarly to the corresponding definition of matrices X_b and Z_b that is $X_c^s = [1, (\boldsymbol{\tau}_s^{(b)})^T]_{1 \le s \le K_b}$, $Z_c = Z_{K_c}\Omega_{K_c}^{-1/2}$ with $Z_{K_c} = [\|\boldsymbol{\tau}_s^{(b)} -$

$\boldsymbol{\tau}_t^{(c)}\|^2 \log \|\boldsymbol{\tau}_s^{(b)} - \boldsymbol{\tau}_t^{(c)}\|]_{1\leq s\leq K_b, 1\leq t\leq K_c}$ and $\Omega_{K_c} = [\|\boldsymbol{\tau}_t^{(c)} - \boldsymbol{\tau}_s^{(c)}\|^2 \log \|\boldsymbol{\tau}_t^{(c)} - \boldsymbol{\tau}_s^{(c)}\|]_{1\leq s,t\leq K_c}$ where the $\boldsymbol{x}$ covariates are replaced by knots $\boldsymbol{\tau}^{(b)}$ and the knots are replaced with subknots $\boldsymbol{\tau}^{(c)}$. The model is completed by adding to (11) the hierarchical structure

$$\Sigma_b = \text{diag}[\exp(X_c\gamma + Z_c c)], \quad c \sim N(0, \sigma_c^2 I_{Kc}).$$

Estimation can now be carried out analogously to above. The knots can be selected with *clara* algorithm described in Kaufman & Rousseeuw (1990). This procedure is implemented in the R package "cluster".

3.2 Simulations and comparisons with other surface fitting methods

For comparison with Crainiceanu, Ruppert & Carroll (2006) and Lang & Brezger (2004) we consider the following regression function with moderate spatial variability

$$m_3(x_1, x_2) = x_1 \sin(4\pi x_2),$$

with x_1 and x_2 independently uniform distributed on $[0, 1]$. We used $n = 300$, $\sigma = 1/4\text{range}(m_3)$ and equally-spaced 12×12 and 5×5 knot grids for $\tau_i^{(b)}$ and $\tau_j^{(c)}$, respectively. Figure 11 displays the resulting fit for one simulation, using our approach. For comparison the true function and the non-adaptive fit are presented in Figures 9 and 10, respectively. Figure 12 visualizes the estimated variance of random effects. We simulated 500 datasets to compare log(MSE) of our estimator with values reported in Crainiceanu, Ruppert & Carroll (2006) and Lang & Brezger (2004). Our simulations provide a median of log(MSE) of -3.79 with an interquartile range [-4.17,-3.80] and a range [-4.96,-2.27]. This is comparable with the results in Crainiceanu, Ruppert & Carroll (2006) (median -3.67, interquartile range [-3.80,-3.53] and a range [-4.21,-3.13]) which outperform the findings of Lang & Brezger

(2004). The average coverage probability of the 95% confidence intervals results to 94.31%. The smoothed coverage probabilities are displayed on Figure 13. Similarly to the Crainiceanu, Ruppert & Carroll (2006), the coverage probability is lowest for $x_1 \in [0.2, 0.5]$. This is explained by the low signal-to-noise ratio in this region.

4 Non-normal response model

4.1 Hierarchical modelling

The technique is now extended to non normal response models by considering the following generalized linear hierarchical mixed model

$$\begin{aligned} E(Y|b,c) &= \mu^{b,c} = h(X_b\beta + Z_b b),\ \mathrm{Var}(Y|b,c) = \phi v(\mu^{b,c}), \\ b|c &\sim N(0,\Sigma_b),\ \Sigma_b = \mathrm{diag}[\exp(X_c\gamma + Z_c c)], \\ c &\sim N(0,\sigma_c^2 I_{K_c}), \end{aligned}$$

with function $h(.)$ as the inverse of link function $\tilde{g}(.)$, $v(.)$ as some specified variance function and ϕ as dispersion parameter. We follow Breslow & Clayton (1993) and estimate the parameters from the quasi-likelihood

$$\exp[ql(\beta,\gamma,\sigma_c^2)] = (2\pi)^{-\frac{(K_b+K_c)}{2}}\sigma_c^{-K_c}\int_{R^{K_b}}\int_{R^{K_c}}\exp[-k_1(b,c)]db\,dc, \qquad (12)$$

with

$$k_1(b,c) = \frac{1}{2\phi}\sum q_i(y_i,\mu_i^{b,c}) + \frac{1}{2}b^T\Sigma_b^{-1}b + \frac{1}{2}\log|\Sigma_b| + \frac{1}{2\sigma_c^2}c^Tc$$

and

$$q_i(y,\mu) = -2\int_y^\mu \frac{y-t}{v(t)}dt,$$

as deviance measure of the fit. Assuming that conditionally on b and c the observations are drawn from the exponential family $Y|b,c \sim \exp[(y\vartheta(x) - b(\vartheta(x)))/\phi +$

$c(y,\phi)]$, the quasi-likelihood (12) represents the true likelihood of the data. Using Laplace's method for approximation of the integral over b, one gets

$$\exp[ql(\beta,\gamma,\sigma_c^2)] \;\approx\; (2\pi)^{-\frac{K_c}{2}}\sigma_c^{-K_c}\int_{R^{K_c}}\exp[-k_2(c)]dc, \tag{13}$$

with

$$k_2(c)=\frac{1}{2}\log|I_n+Z_b^TWZ_b\Sigma_b|+\frac{1}{2\phi}\sum q_i(y_i,\mu_i^{b,c})+\frac{1}{2}\hat{b}^T\Sigma_b^{-1}\hat{b}+\frac{1}{2\sigma_c^2}c^Tc,$$

where $\hat{b}$ is the solution to

$$\frac{\partial k_1(b,c)}{\partial b}=-Z_b^TW\mathrm{diag}[\tilde{g}'(\mu^{b,c})](Y-\mu^{b,c})+\Sigma_b^{-1}b=0,$$

with W as the $n\times n$ diagonal matrix of GLM iterated weights with diagonal elements $w_i=(\phi v(\mu_i^{b,c})[\tilde{g}'(\mu_i^{b,c})]^2)^{-1}$, using the simplifying assumption that the iterative weights w_i vary only slowly (or not at all) with the of mean.

Substituting the current estimate $\hat{b}$, say, into (13) and replacing the deviance $\sum q_i(y_i,\mu_i^{b,c})$ in $k_2(.)$ by the Pearson chi-squared statistic $\sum(y_i-\mu_i^{b,c})^2/v_i(\mu_i^{b,c})$ result to

$$\exp[ql(\beta,\gamma,\sigma_c^2)] \;\approx\; (2\pi)^{-\frac{K_c}{2}}\sigma_c^{-K_c}|W|^{-1/2}\int_{R^{K_c}}\exp[-k_3(c)]dc,$$

with

$$k_3(c)=\frac{1}{2}\log|V|+\frac{c^Tc}{2\sigma_c^2}+(U-X_b\beta)^TV^{-1}(U-X_b\beta),$$

where $V=W^{-1}+Z_b\Sigma_bZ_b^T$ and $U=X_b\beta+Z_b\hat{b}+\mathrm{diag}[\tilde{g}'(\mu^{b,c})](Y-\mu^{b,c})$. Applying again Laplace's method, we end up with the following quasi-log-likelihood for the remaining parameters

$$\begin{aligned}-2l(\beta,\gamma,\sigma_c^2) &\approx K_c\log\sigma_c^2+\log|V|+\log|k_3^{cc}|\\ &+ \hat{c}^T\hat{c}/\sigma_c^2+(U-X_b\beta)^TV^{-1}(U-X_b\beta),\end{aligned}$$

with $k_3^{cc} = \partial^2 k_3(c)/\partial c \partial c^T$. In complete analogy to Section 2 the estimation of parameter $\theta = (\gamma^T, c^T)^T$ can be carried out from the score equation

$$\frac{\partial k_3(\hat{\theta})}{\partial \theta} = -\frac{1}{2} W_c^T \Sigma_b^{-1} \left\{ \hat{b}^2 - w_{df} \sigma_b^2 \right\} + D_c \theta / \sigma_c^2 = 0. \tag{14}$$

Numerically this procedure can be implemented by iterating between estimation of $\hat{\theta}$ and thus $\hat{\Sigma}_b$ from (14) and calls of any generalized linear mixed models software.

4.2 Simulations for non-normal response

We consider the following model for binomial data $Y_i \sim B(n_i, \pi_i)$ with canonical link

$$\begin{aligned} \text{logit}[E(Y_i|x_i)/n_i] &= \log\left(\frac{\pi_i}{1-\pi_i}\right) = X_b^i \beta + Z_b^i b,\ i = 1, ..., n, \\ b|c &\sim N(0, \Sigma_b),\ \Sigma_b = \text{diag}[\exp(X_c \gamma + Z_c c)], \\ c &\sim N(0, \sigma_c^2 I_{K_c}). \end{aligned}$$

The diagonal elements of the iterative weights in matrix W for this model equal $w_i = 1/n_i \pi_i (1 - \pi_i)$. We simulated data with probabilities $\pi = \text{logit}^{-1}(m_2(x))$, where function $m_2(.)$ is the same as in Section 2. Figure 14 represents exemplary the fit for the grouped data with $n_i = 5$ and $n = 1000$ (bold). Figure 15 shows the fit for $n = 5000$ binary data (bold). For comparison the fit with global smoothing parameter is also presented (dashed). The benefit of local adaptivity is obvious.

There is no available adaptive smoothing method implemented for binary responses, so that we can not compare our routine to other approaches. However, the BARS procedure of DiMatteo, Genovese & Kass (2001) allows to fit Poisson responses. We performed a number of simulations to compare the performance of our routine with the BARS implemented for this setting. We simulated $n = 800$ Poisson distributed data with means $\exp[m_1(x)]$, where $m_1(.)$ is the same as in Section 2.5 with $j = 4$.

We estimated the data with our approach using $K_b = 60$ and $K_c = 10$ and with the BARS procedure, letting MCMC chain run for 10000 iterations with a burn-in period of 2000. Figure 16 displays estimates based on our adaptive approach (bold) and BARS (dashed). The AMSE of the five fits based on our approach equals 0.010672, while the AMSE of BARS based fits is 0.020547. We did not perform a more extensive simulation study, since a single BARS fit required in general more than 4 hours estimation time on an up-to-date computer. For comparison, estimation with our function `asp` was carried out within a minute. We experimented with other mean functions and sample sizes as well, overall obtaining similar results.

4.3 Example

For demonstrational purposes we apply the above spatially adaptive smoothing technique to a dataset on the absenteeism of workers of a company in Germany. Parts of the data have been analysed before in Kauermann & Ortlieb (2004) with a different focus though. We consider absenteeism spells and model the probability of returning to work after a sick leave. Denoting the duration of such a leave by d, we model the discrete hazard rate

$$P(d = t|d \geq t) = h(t), \tag{15}$$

where $t = 1, 2, ...$ The duration is thereby measured in days and the event of interest is the recovery which allows workers to return to work. If the worker has reported sick on one day, say Tuesday, but returns to work on a consecutive working day thereafter, we count this as event and the duration is the number of working days the worker has been absent. If, in contrast, the last days of absenteeism and the first day of returning to work are not consecutive working days, we consider the duration as censored observation, and d gives the number of days of absenteeism.

To make this more explicit, assume that a worker reports sick on Friday but returns to work the Monday after. It is unclear when the worker actually recovered, either already Friday, Saturday or Sunday. It is however known, that the worker was at least sick on one day and the observation is therefore $d = 1$ with censoring indicated. Let now δ denote the censoring indicator which is either zero, for censoring, or 1, otherwise. For each absence spell we transform d to the binary variables $y_1, ..., y_d$ with $y_l = 0$ for $l < d$ and $y_d = \delta$. The hazard function is then the probability $P(y_t = 1|y_l = 0, l < t)$. We concentrate on short term absenteeism spells truncated at $d = 10$ and take longer spells as censored observations. Besides the explicit duration time we allow the hazard function to depend on calendar time c as well, where c is the first day of the absenteeism spell of the worker. The final model is then

$$\text{logit}P(d = t|d \geq t, c) = m(t, c), \tag{16}$$

which is fitted in a local adaptive way below.

The data were collected in company in South Germany and we analyse the data of about 370 employees. Not all of them were employed at the same time with the observation period ranging from 1981 to 1998. On average, about 3/4 of the employees reported sick at least once per calender year. We assume that the durations of different sick leaves of the same worker are independent and even though it might be argued whether this is an appropriate assumption, for sake of simplicity we leave this issue aside for now. Figure 17 and 18 show the fit of the model (16) using non-adaptive and adaptive smoothing, respectively. Both fits were carried out using 14 knots for each dimension and low-rank thin spline basis as defined in Section 3.1. The variance structure for the adaptive fit was modelled with 10 knots for each dimension. The differences in the plots are quite obvious. Both fits expose

a bump at 1992 and 1993 and day 3, which becomes even more peaked for the spatially adaptive fit. Beyond this peak, particularly for longer absenteeism time, the non-adaptive fit is quite wiggled while the adaptive approach selects a smooth, flat behaviour. The latter fit looks preferable and once more demonstrates the benefits of spatial adaptivity.

The peak at year 1993 and duration time at day 3 allows for an interesting economic interpretation. In 1992/93 the company went through a major downsizing process with more than 50% of the workers being dismissed. While this economic situation has hardly any effect on the hazard function for days $d \geq 5$, it does affect the hazard rate for short absenteeism times, particularly for $d = 3$. Due to the German law, workers reported sick for more than 3 consecutive working days have to provide a medical certificate at the latest at the third day of their sick leave, while for shorter periods no special attestation is required. Apparently, during the downsizing period the duration of sick leaves is clearly shorter with more employees returning after 3 days. This provides indication that economic critical conditions of a company have a direct influence on the absenteeism of employees. Looking further in the data it can be seen that it are mainly employees who are being dismissed who tend to change their absenteeism behaviour (see also Kauermann & Ortlieb, 2004), while Figure 18 shows how this is changed. Moreover, the locally adaptive smoothing exposes the peak more clearly without overfitting the remaining regions and therefore justifies the additional modelling effort.

5 Conclusion

We demonstrated how local adaptive smoothing can be easily carried out by formulating penalties on spline coefficient as hierarchical mixed model. The major

contribution was to show how simple Laplace approximation of the marginal likelihood allows to fit such models relatively easy and fast without MCMC methods. For reasonably sized data sets our routine needs seconds while any MCMC routine needs minutes. In addition, small changes to the model, such as adding a covariate can be handled very easily in our implementation, whereas it may take hours, days, or even weeks with MCMC software. Another reason why having a fast and accurate procedure is useful is that in many situations the smoothing procedure needs to be applied repeatedly. One trivial example is when doing simulations. Moreover our approach can easily be extended to more general settings like spatial smoothing or generalized response models.

A R Package "AdaptFit"

To implement our approach we developed an R package. We took advantage of the R package "SemiPar", written by M.P. Wand to accompany the book Ruppert, Wand & Carroll (2003). The function `spm` of this package performs scatterplot, spatial and generalized (binomial and poisson) smoothing using the (generalized) mixed models representation of penalized splines. This function handles additive models as well. To perform adaptive smoothing we had to integrate the Fisher scoring procedure (9) for θ with updates of the remaining parameters by subsequent calls of function `spm`. The current version of our package "AdaptFit" with the function `asp` is available at http://cran.r-project.org. In general, the usage of `asp` is similar to that of function `spm`. For example, estimation of the function $m_1(x)$ described in Section 2.5 can be performed with

```
> x <- 1:400/400
> mu <- sqrt(x*(1-x))*sin((2*pi*(1+2^((9-4*6)/5)))/(x+2^((9-4*6)/5)))
```

```
> y <- mu+0.2*rnorm(400)
> kn <- default.knots(x,80)
> kn.var <- default.knots(kn,20)
> y.fit <- asp(y~f(x,knots=kn,var.knot=kn.var))
> plot(y.fit)
```

Switching between maximum likelihood and restricted maximum likelihood estimation can be done by specifying `spar.method="ML"`. Other examples are provided within the package. In the additive model case `asp` also allows to fit some components of the model non-adaptively.

References

Baladandayuthapani, V., Mallick, B., and Carroll, R. (2005). Spatially adaptive Bayesian penalized regression splines (P-splines). *Journal of Computaional and Graphical Statistics* **14**, 378–394.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association.* **88**, 9–25.

Crainiceanu, C., Ruppert, D., and Carroll, R. (2006). Spatially adaptive Bayesian P-splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, (to appear).

Crainiceanu, C., Ruppert, D., and Wand, M. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of statistical software* **14(14)**.

DiMatteo, I., Genovese, R., and Kass, R. (2001). Bayesian curve-fitting with free-knots splines. *Biometrika* **88**, 1055–1071.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Stat. Science 11*(2), 89–121.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B* **57**, 371–394.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modelling. *Technometrics* **31**, 3–39.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association.* **72**, 320–338.

Herrmann, E. (1997). Local bandwidth choice in kernel regression estimation. *Journal of Computational and Graphical Statistics* **6**, 35–54.

Kass, R. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayesian models). *Journal of the American Statistical Association.* **84**, 717–726.

Kauermann, G. (2004). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planing and Inference* **127**, 53–69.

Kauermann, G. and Ortlieb, R. (2004). Temporal pattern in the number of staff on sick leave: the effect of downsizing. *Journal of the Royal Statistical Society, Series C* **53**, 353–367.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Laird, N. and Louis, T. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *Journal of the American Statistical Association.* **82**, 739–757.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Association* **92**, 107–116.

Morris, C. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *Journal of the American Statistical Association.* **78**, 47–65.

Ngo, L. and Wand, M. (2004). Smoothing with mixed model software. *Journal of statistical software* **9(1)**.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (c/r: P519-527). *Statistical Science* **1**, 502–518.

Pintore, A., Speckman, P., and Holmes, C. C. (2005). Spatially adaptive smoothing splines. *Biometrika*, (to appear).

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

Ruppert, D. and Carroll, R. (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* **42**, 205–224.

Ruppert, R., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press.

Searle, S., Casella, G., , and McCulloch, C. (1992). *Variance Components*. Wiley.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics* **18**, 223–249.

Wood, S., Jiang, W., and Tanner, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* **89**, 513–528.

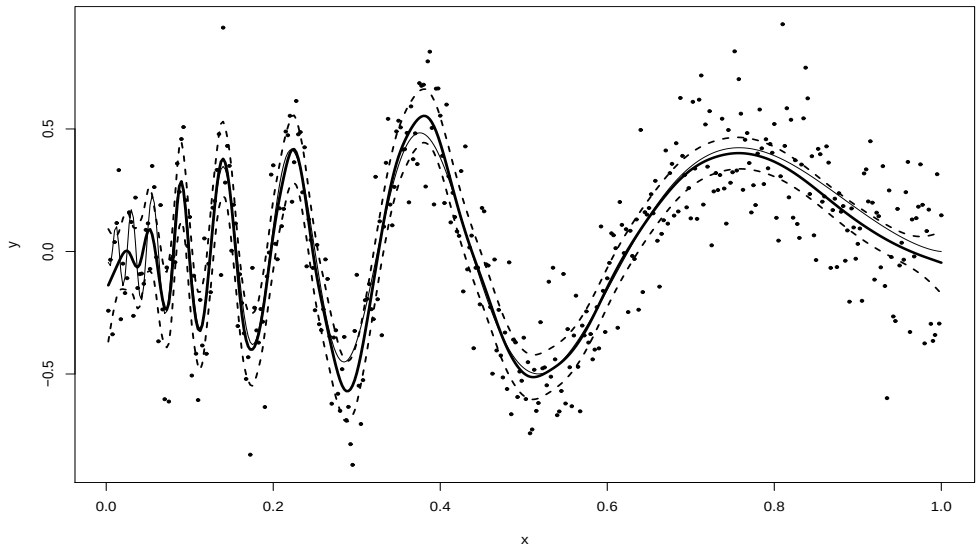

Figure 1: Estimated regression function $m_1(x)$ (bold) with confidence intervals (dashed) and true function.

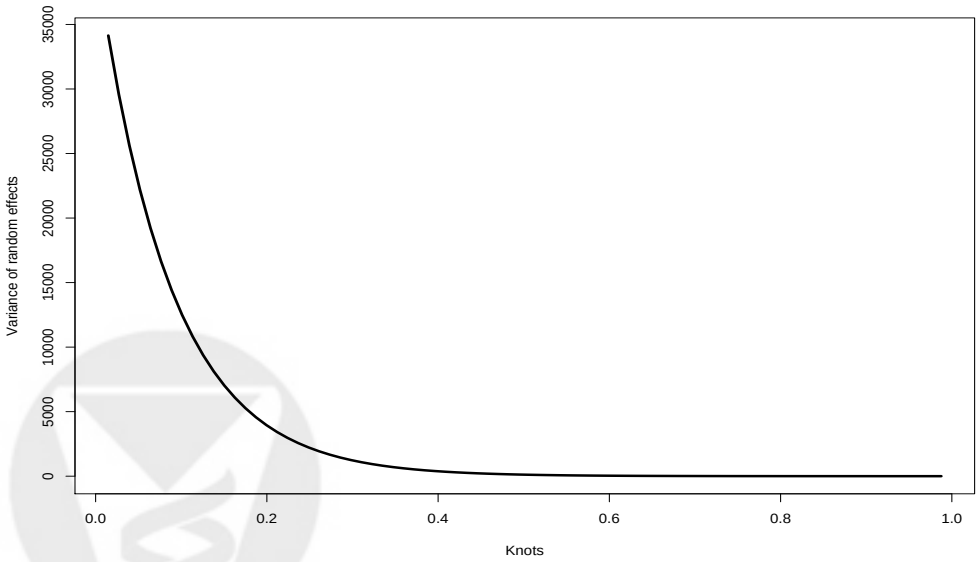

Figure 2: Estimated variance of random effects for the regression function $m_1(x)$.

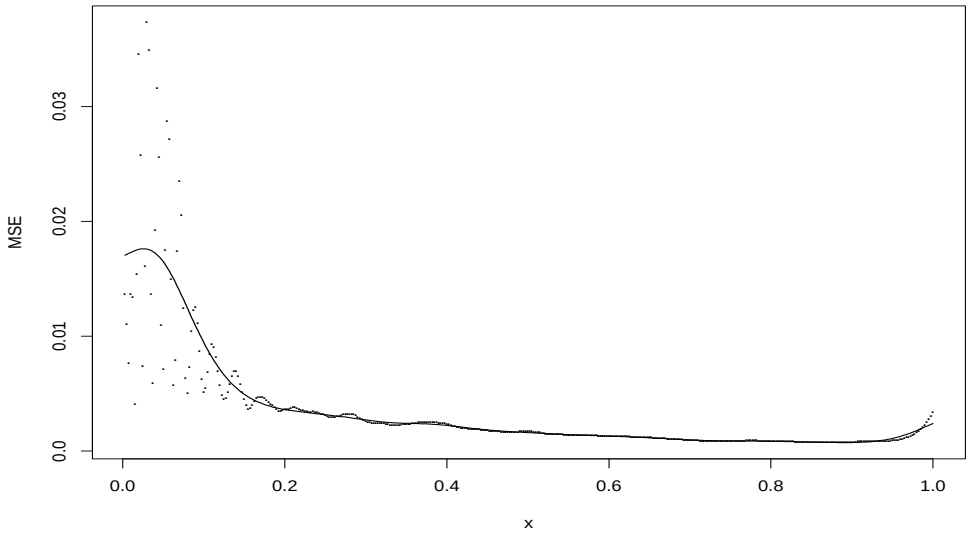

Figure 3: Pointwise MSE for 500 simulated datasets with function $m_1(x)$. Solid line shows a smoother of the points.

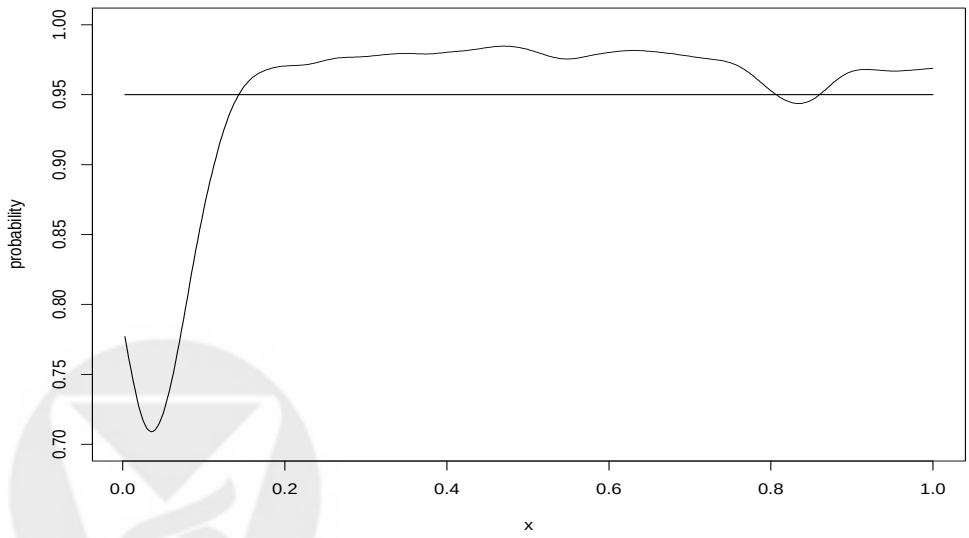

Figure 4: Smoothed pointwise coverage probabilities of 95% confidence intervals for 500 simulated datasets with function $m_1(x)$.

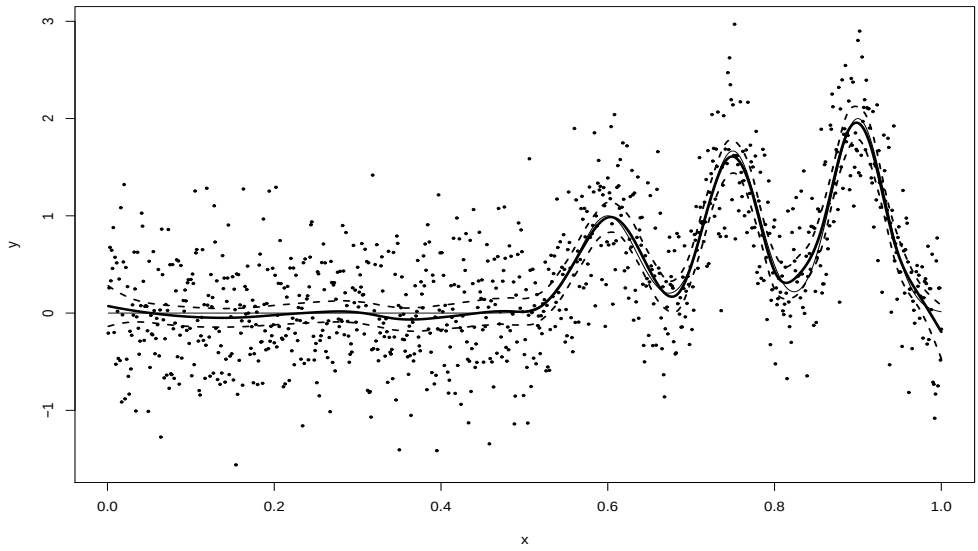

Figure 5: Estimated regression function $m_2(x)$ (bold) with confidence intervals (dashed) and true function.

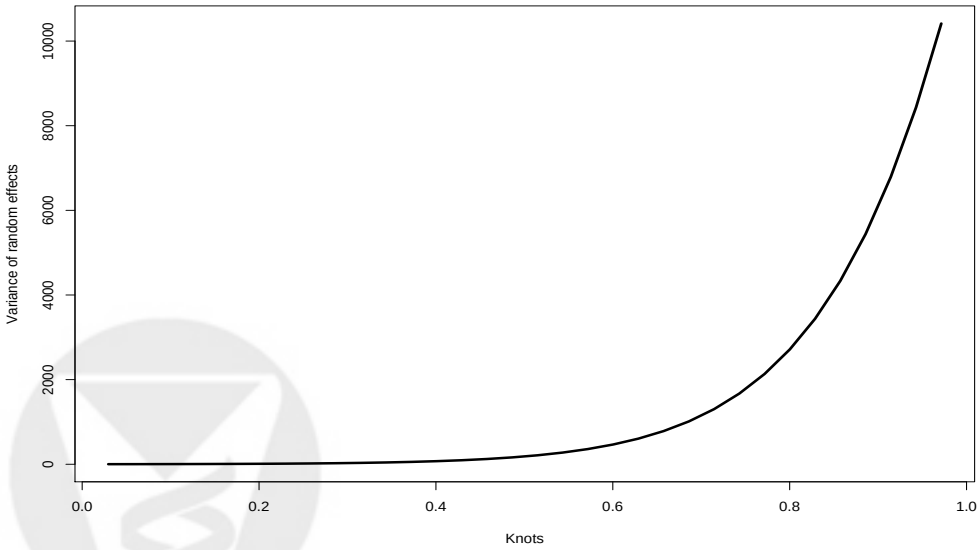

Figure 6: Estimated variance of random effects for the regression function $m_2(x)$.

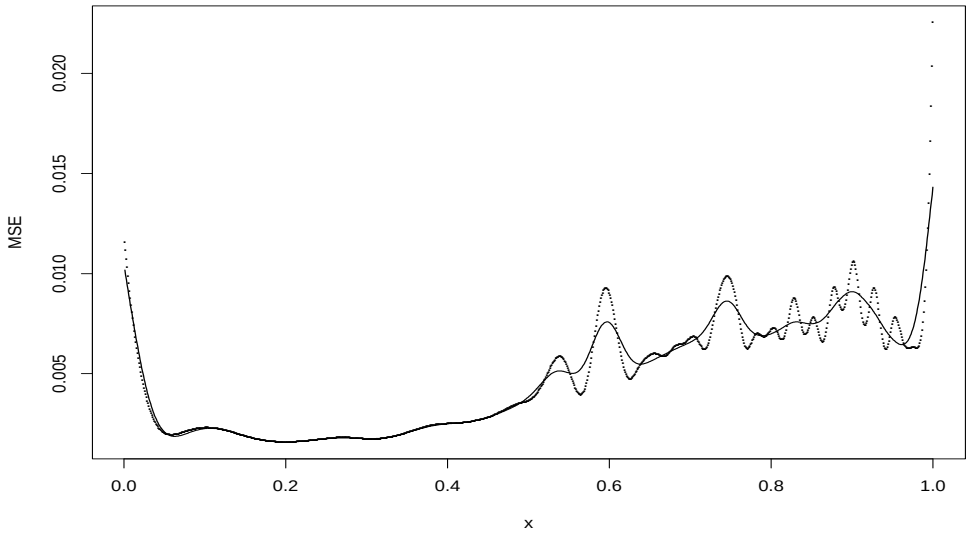

Figure 7: Pointwise MSE for 500 simulated datasets with function $m_2(x)$. Solid line shows a smoother of the points.

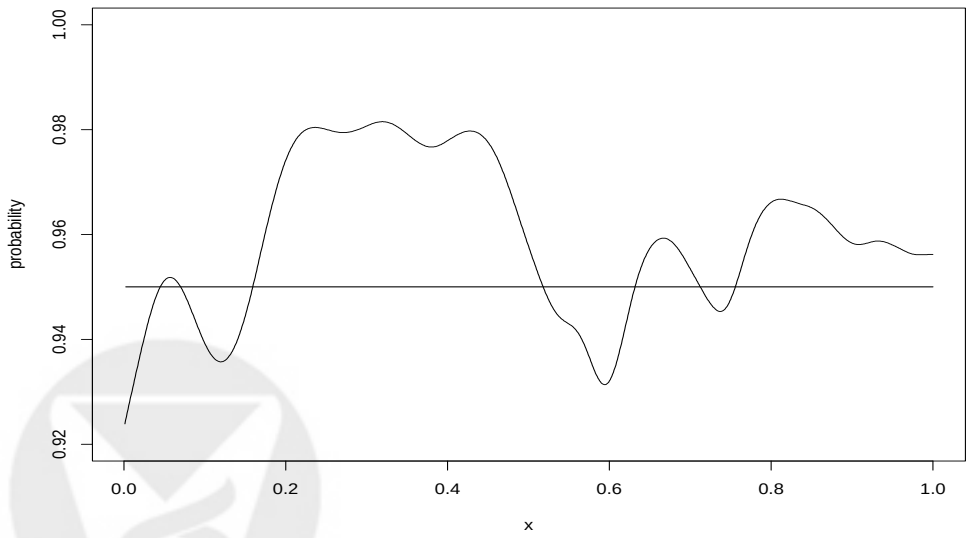

Figure 8: Smoothed pointwise coverage probabilities of 95% confidence intervals for 500 simulated datasets with function $m_2(x)$.

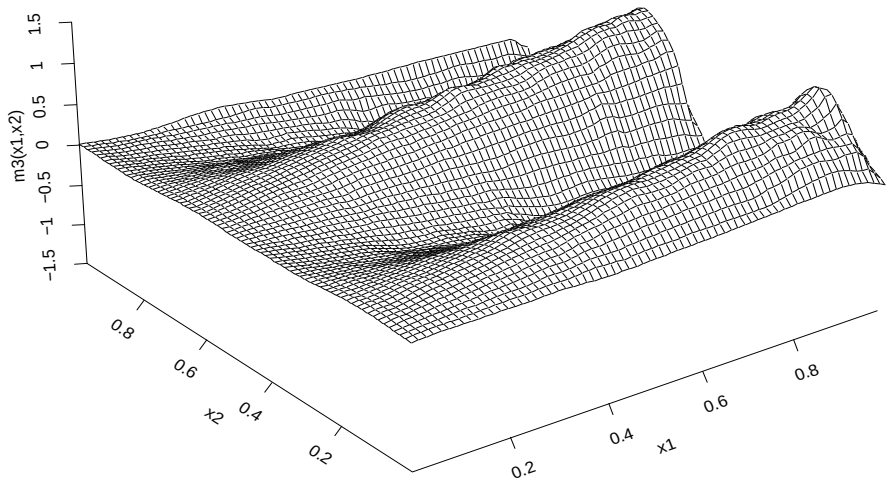

Figure 9: Regression function $m_3(x_1, x_2)$

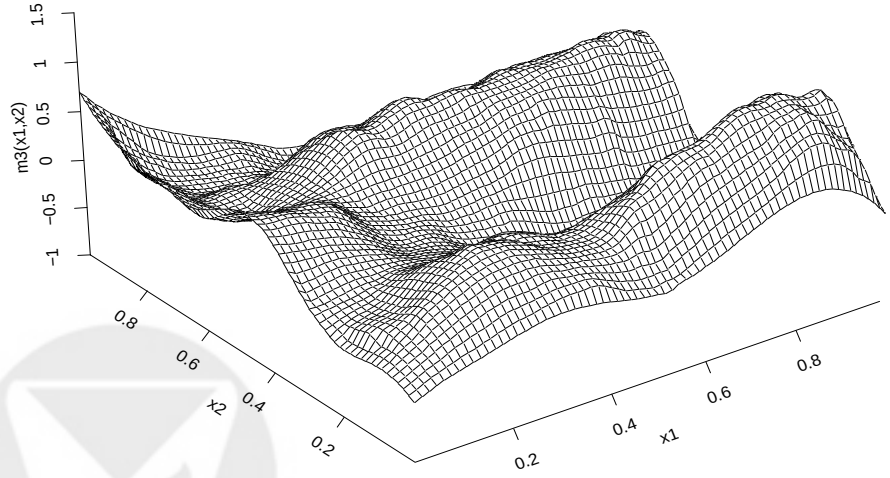

Figure 10: Estimated regression function $m_3(x_1, x_2)$ with global smoothing parameter.

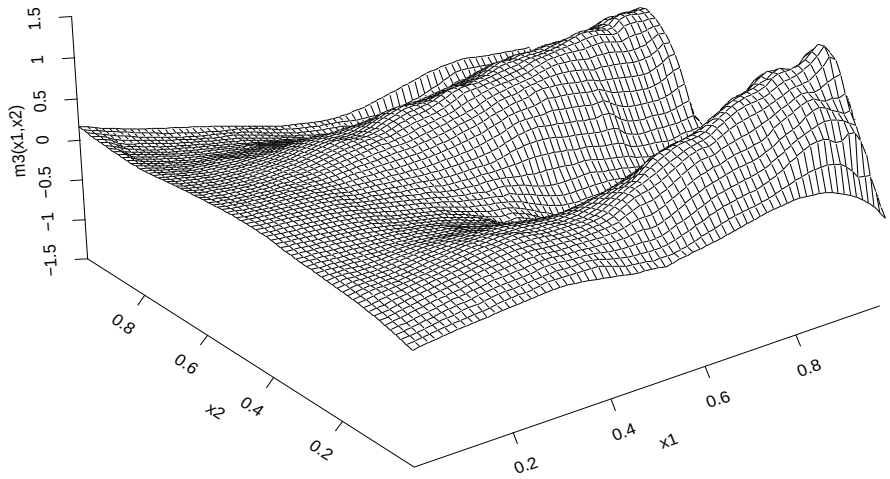

Figure 11: Estimated regression function $m_3(x_1, x_2)$ with adaptive penalty.

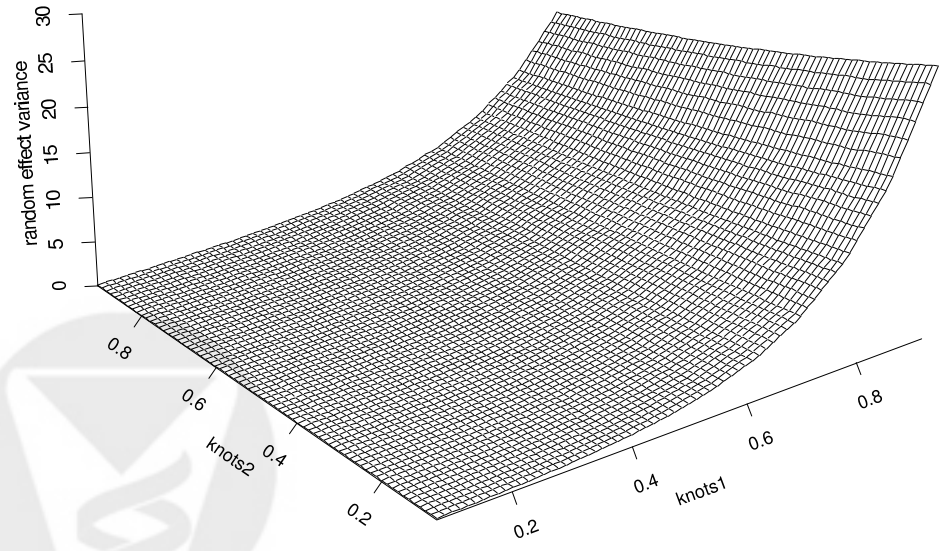

Figure 12: Estimated variance of random effects of the regression function $m_3(x_1, x_2)$.

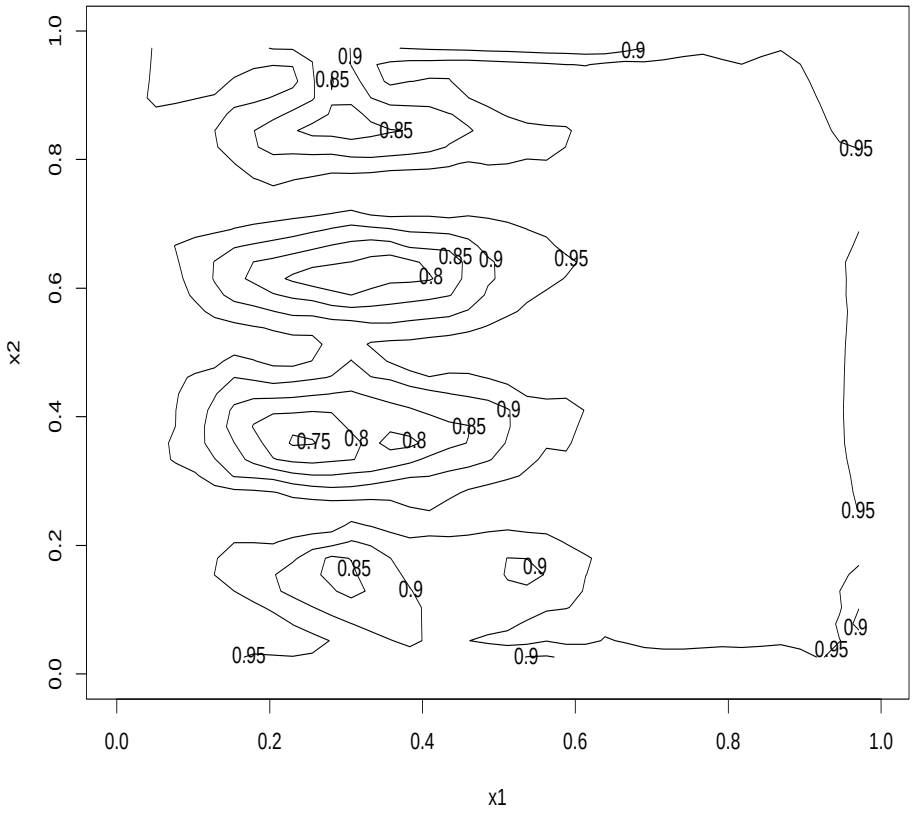

Figure 13: Smoothed coverage probability of 95% confidence intervals for 500 simulated datasets with function $m_3(x)$.

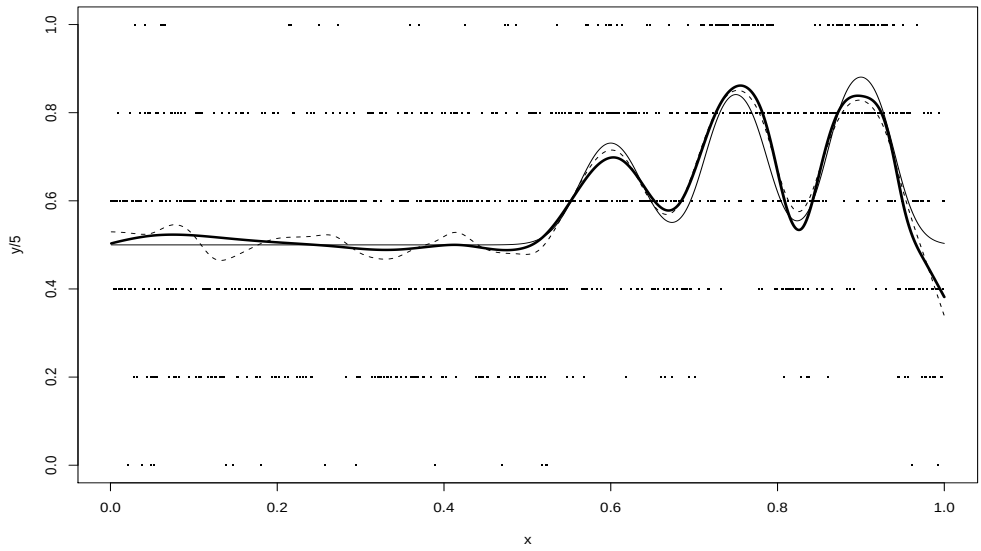

Figure 14: Estimated regression function $\pi = \text{logit}^{-1}(m_2(x))$ with adaptive penalty (bold), with global smoothing parameter (dashed) and true function for 1000 grouped binomial data ($n_i = 5$).

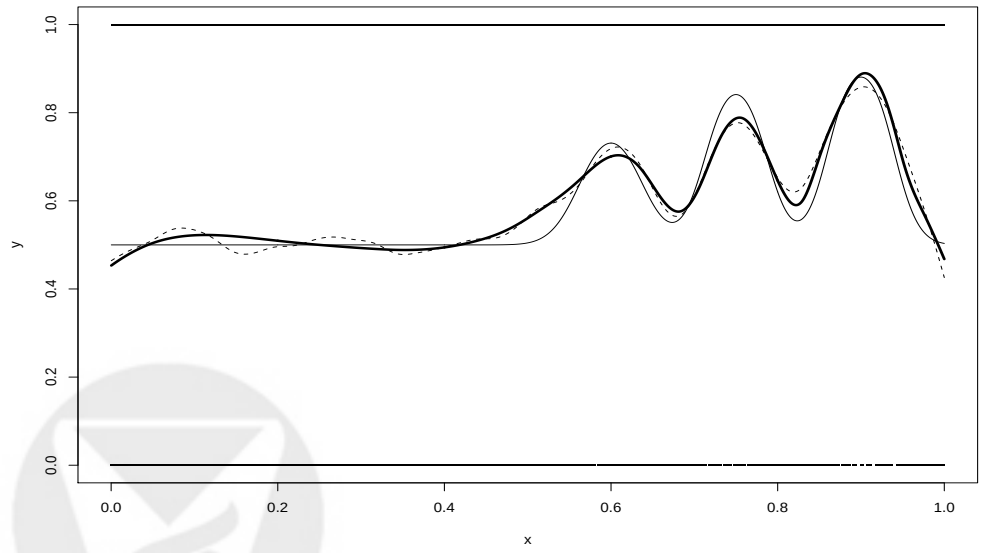

Figure 15: Estimated regression function $\pi = \text{logit}^{-1}(m_2(x))$ with adaptive penalty (bold), with global smoothing parameter (dashed) and true function for 5000 binary data.

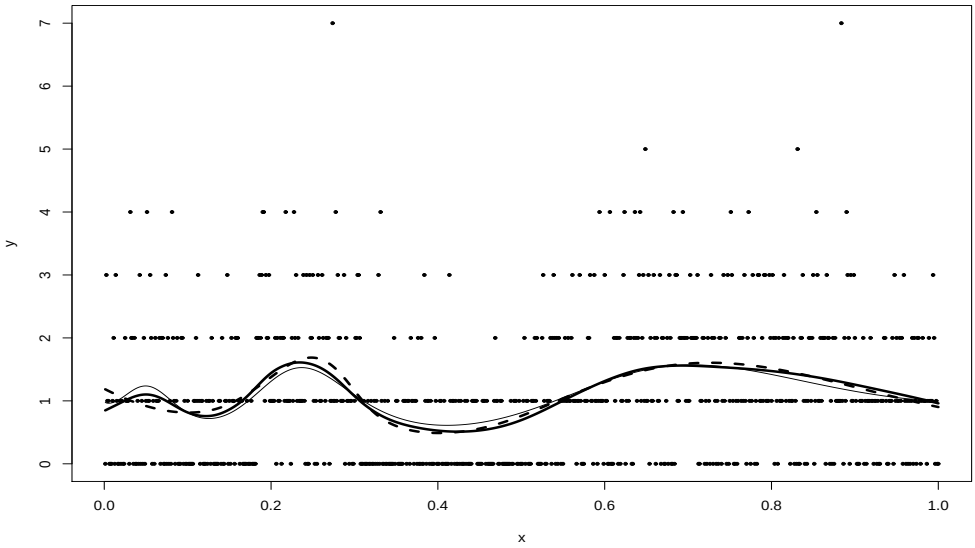

Figure 16: Estimated regression function $\exp[m_1(x)]$ based on our adaptive approach (bold) and BARS (dashed).

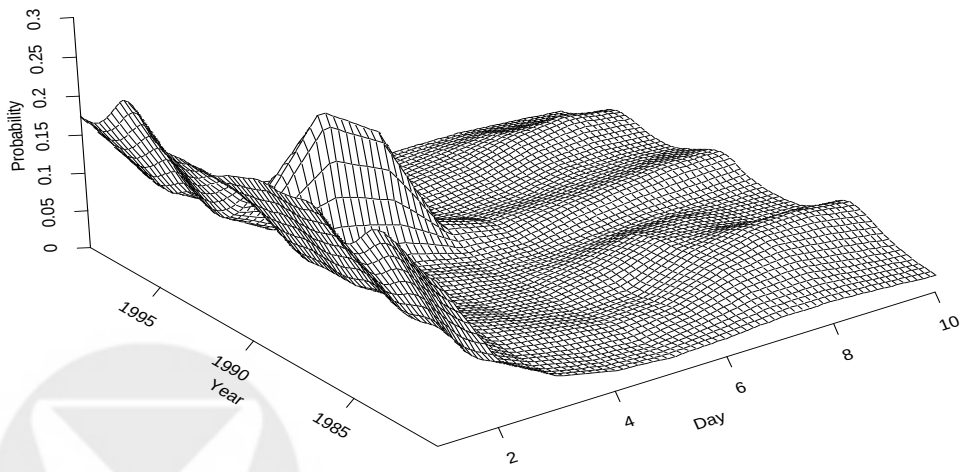

Figure 17: Estimated regression function $P(d = t|d \geq t, c) = \text{logit}^{-1}(m(t, c))$ with global smoothing parameter.

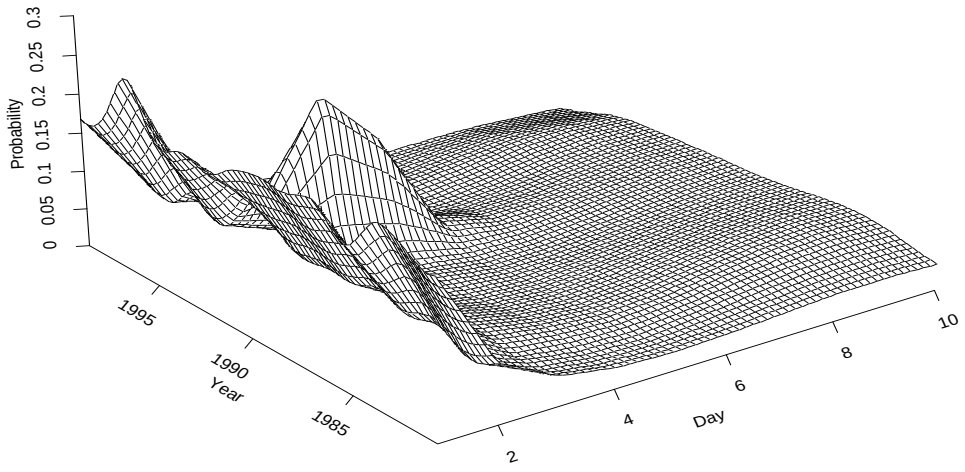

Figure 18: Estimated regression function $P(d = t|d \geq t, c) = \text{logit}^{-1}(m(t, c))$ with adaptive penalty.